TACKLING BACKGROUND DIFFERENTLY

Jetley S., Romera-Paredes B., Torr P. - University of Oxford
{sjetley, bernard, phst}@robots.ox.ac.uk

ICVSS 2015
Sicily ~ 12-18 July
International Computer Vision Summer School

Abstract

Practical vision systems need to identify objects as belonging to one of n different classes of interest while distinguishing them from the remaining image content often clumped together as 'the background class'. Background does not have one consistent definition, yet highly successful approaches based on CNNs[1, 2], treat it as a regular object category and attempt to learn its appearance. This is counter-intuitive.
We propose a modified deepnet architecture for tackling background class as 'something other than the object categories of interest.' We filter out the background samples using a threshold that is learnt via end-to-end training. Separating the background samples allows us to define an unambiguous output embedding space grounded completely in the object classes of interest, which has benefits for the task of zero-shot recognition.

Motivation

Issues with the traditional way of handling 'background' in a deepnet:

- Learning **non-similar visual patterns** as one single background class is counter-intuitive.
- Unique priors based on **natural language semantics, attributes are not available** for background class.

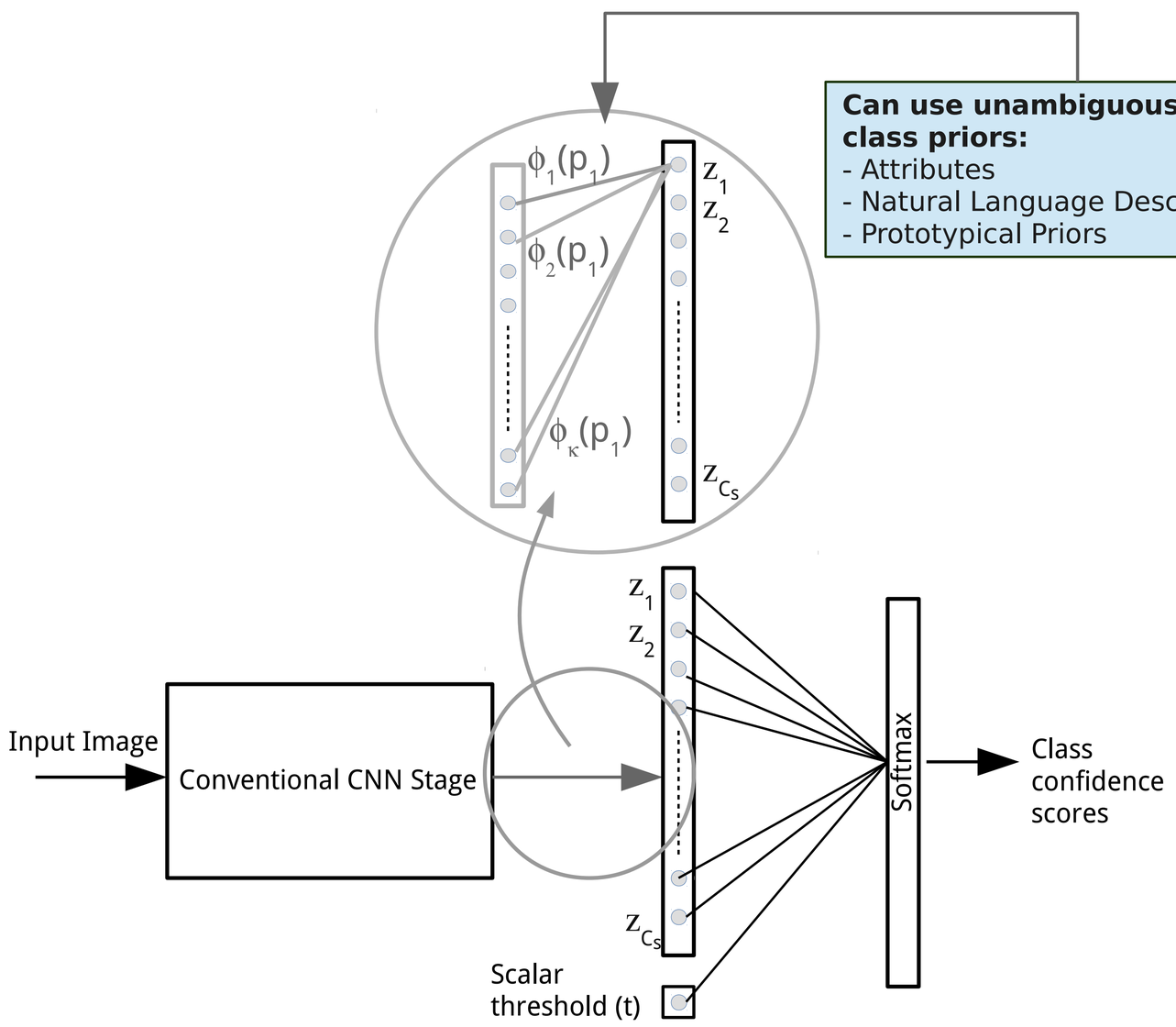
Proposed Approach

The aim is to identify samples of the background class using a threshold T (fixed or learned), such that:

$$x = \begin{cases} \text{background}, & \text{if } s_c < T \; \forall \; c \in \{1, \ldots, C\} \\ \text{object}, & \text{otherwise} \end{cases} \quad (1)$$

, where x is the input sample and s_c is the output activation score (before the softmax operation) for class c and C is the total number of object classes (ref. Fig.1).

Proposed Architecture



Benefits

- As aligned with our natural intuition, background is treated as a visual concept that is dissimilar to any of the object categories of interest i.e. for which **the class activation score for any and every class is <** T.
- Allows **end-to-end learning** of background threshold.
- **Unambiguous class priors** ϕ based on natural language semantics or visual attributes can be leveraged for zero-shot recognition (Ref. Fig.1).

Dataset

German Traffic Sign Dataset (43 classes):
- **Training**: 39209 object samples & 1410 background samples.
- **Val & Test**: 6316 object samples & 450 background samples; 6315 object samples & 525 background samples.

Results

Description	Test Acc.(%)
Background as object	95.77
Fixed-T	95.19
Learned-T	95.68

References

[1] Ren, Shaoqing, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." arXiv:1506.01497 (2015).
[2] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." arXiv:1411.4038 (2014).
[3] Jetley, Saumya, et al. "Prototypical Priors: From Improving Classification to Zero-Shot Learning." BMVC. 2015.